Computational Statistics: Assignment 1

Akos Engelmann, Gergely Paradi, Fabian Gallyas, Ipek Cakin, Simon Jasansky

2023-03-05

1 Define the assumptions that you would like to make.

First it must be noted that we test two different approaches. Both rely on the same assumptions, which are:

1. There is a causal relationship between the variables
2. The causal relationship(s) (i.e. the graph) are acyclic and with no self-loops
3. The model is an additive noise model, i.e. the variance of a node is at least as big as the variance of the one that it is directed from, plus some additional noise ϵ. This noise ϵ is independent and identically distributed (i.e. noise of all variables comes from the same distribution) with a normal distribution with mean 0 and standard deviation σ.

2 Develop a procedure to estimate the DAG from the data using your assumptions.

We tested two methods to create DAGs.

2.1 First Method: Brute force regression

The first approach is a simple one. In the first step, we took a simpler but insightful approach to see the data clearly and determine the appropriate approach. For that reason, we took each of our variables one by one as the target variable and regressed them with all the other ones.

Then we checked the significance of these connections by checking the p-values (significant if p-value $<$ 0.05). When the t-test suggested that the predictor variable has a significant effect on the target variable, we established a connection.

After creating a connectivity matrix, we tested the variances between the connected nodes with an F-test and by doing so, we identified the directions with the following method:

- No arrow between two nodes, if there is no significance. (meaning no influence between x and y)
- x -> y if variance of x is less than variance of y. (meaning x influences y)
- x <- y if variance of x is greater than variance of y. (meaning y influences x)

Although it can be desirable to include all of the variables when regressing there are some issues with it as well, which will be discussed in the limitations. We checked for multicollinearity and the value was not significant but also not ideal. This method closely follows the algorithm proposed in Park (2020), with some deviations.

2.2 Second method: sortnregress

For the second approach we used the sortnregress method, which was introduced during the lecture and is based on Reisach, Seiler, and Weichwald (2021). To summarize the approach, with sortnregress the variables are initially ranked by their variances, from lowest to highest variance. As we assume an additive noise model, a variable that is influenced by another variable must have at least the noise of the variable that it is

influenced by, plus some additional noise. Then, each variable is linearly regressed on all the other variables having lower variance than the variable itself.

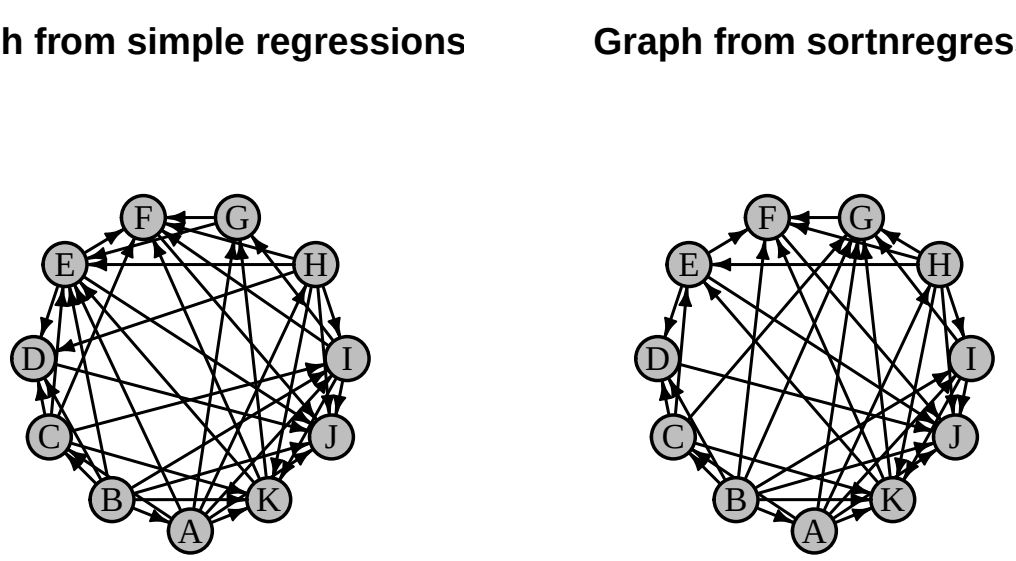
In detail, first a linear regression is performed to get coefficients. After that, a LASSO regression is performed, where the regressors are multiplied by the absolute value of the coefficients of the linear regression, making variables that have only a small effect in the linear regression even smaller in the LASSO regression. The LASSO regression lets coefficients for less important variables go to zero, having the effect of pruning connections between variables in the DAG.

3 Implement your procedure in R or Python and report your DAG

The code used for this assignment can be found at https://github.com/SimonJasansky/comp_stat_A1. Our code for sortnregress closely follows the code provided by Reisach, Seiler, and Weichwald (2021), that is available at https://github.com/Scriddie/Varsortability. For the regressions used in the sortnregress algorithm, the scikit-learn implementation of Linear regression and LASSO regression were used, where the LASSO regression uses least-angle regression (LARS) with the Bayesian information criterion to prune varibles.

3.1 Comparing the two models

In the following figure, you can see the DAGs resulting from the two implemented methods.



As seen in DAG, there results are similar for most nodes and edges in the two methods, with some differences. There are 11 edge differences in total between the two methods, corresponding to 10% in the whole data set (121 possible edges). While 5 nodes which are D,E,F,J and K gave exactly the same results, 1 or 2 edge differences were observed in the other 6 nodes that are A,B,C,G,H and I. Notably, the sortnregress method produced a more sparse graph having less edges than the simple regression, which intuitively makes sense as

there are less regressors, and it also includes a shrinkage penalty to prune edges.

The main differences of the two methods is in the edges coming to the E and G nodes. While in the first approach there were more incoming edges to the E node - the total incoming edge was 6, in the sortnregress it is reduced to 3. On the other hand, while the edges coming to the G node were 3 in the brute force approach, they increased to 6 in the sortnregress approach.

4 Explain what the limits of your procedure are, both statistically and in terms of possible violations of your assumption with real world data

Our DAG procedure assumes that there is a causal relationship between the variables, that the causal relationship is acyclic and has no self-loops, and that the errors follow an additive noise model with uniform, independent errors. Violations of these assumptions can lead to biased estimates of the causal relationships.

As mentioned above, our first approach is to one variable on all other variables. However, if one of the other variables is an intermediate variable, we should not adjust for it in the regression, as it can lead to the Simpson's paradox.

Given identifiability condition could be very restrictive, even if it is very general version of the conditions for linear structural equation models (SEMs).

Additionally, Additive Noise Models (ANMs) statistical consistency of learning non-linear have been not provided yet.

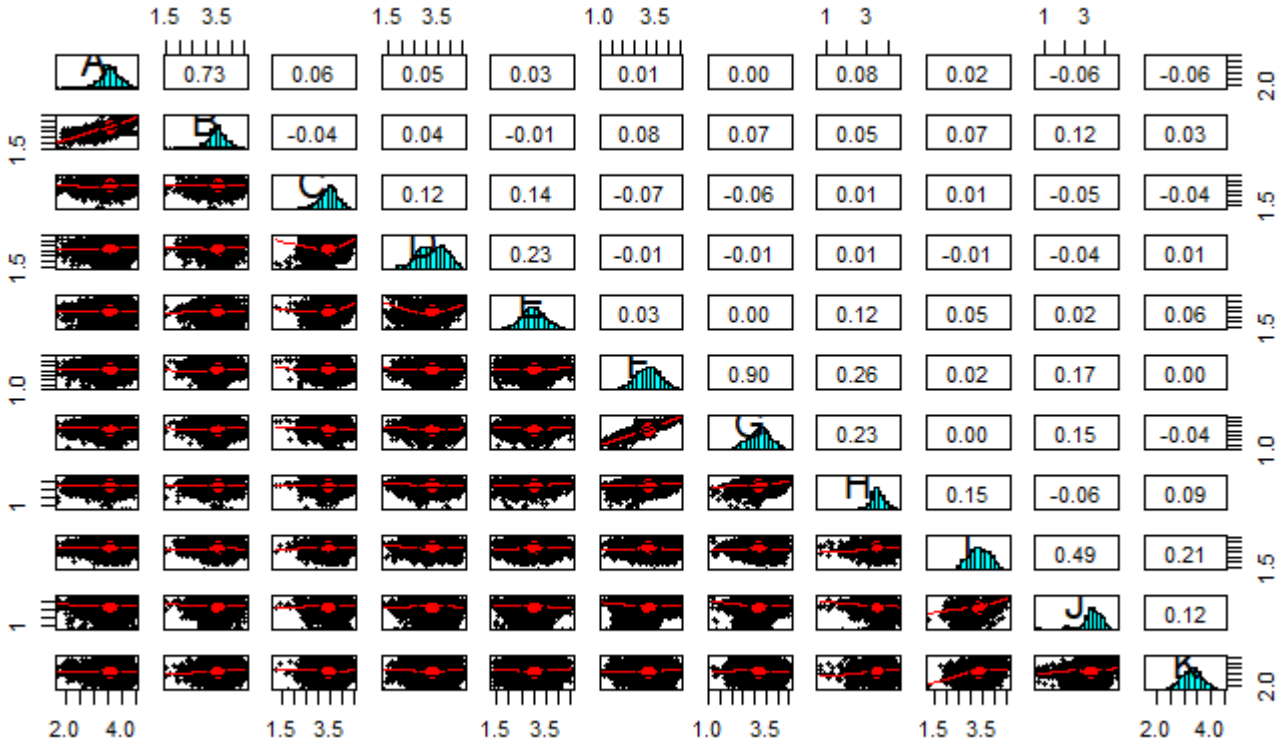
The DAG procedure can identify correlations between variables, but it does not necessarily identify causal relationships. Careful interpretation of the results is necessary to determine whether the relationships identified by the DAG procedure are truly causal.

Furthermore, Real-world data often includes confounding variables, which can complicate the estimation of causal relationships. And also, the assumptions of the DAG procedure may not hold for all types of data, such as data that include discrete variables or time-varying relationships.

Table 1: Variances per variable in decreasing order

vars	variance
B	0.1252109
A	0.1431707
H	0.1454580
C	0.1565856
I	0.1885913
K	0.1906649
G	0.2276126
E	0.2423992
F	0.2834611
D	0.3339794
J	0.4328275

5 Appendix



6 References

Park, Gunwoong. 2020. "Identifiability of Additive Noise Models Using Conditional Variances." *Journal of Machine Learning Research* 21 (75): 1–34. http://jmlr.org/papers/v21/19-664.html.

Reisach, Alexander G., Christof Seiler, and Sebastian Weichwald. 2021. "Beware of the Simulated DAG! Causal Discovery Benchmarks May Be Easy to Game." In *Neural Information Processing Systems*.